# Customer segmentation prediction & Customer clustering

A case study of an automobile dataset from Kaggle

By: Yulong Gong

# CONTENTS

# 1 Business Problem

Predicting consumer segmentation based on past data to choose better marketing strategies to enter a new market.

- Target: customer segmentation.
- Dataset source: https://www.kaggle.com/kaushiksuresh147/customer-segmentation?select=Train.csv

# 2 Data processing and EDA

## 6 out of 11 columns containing missing values

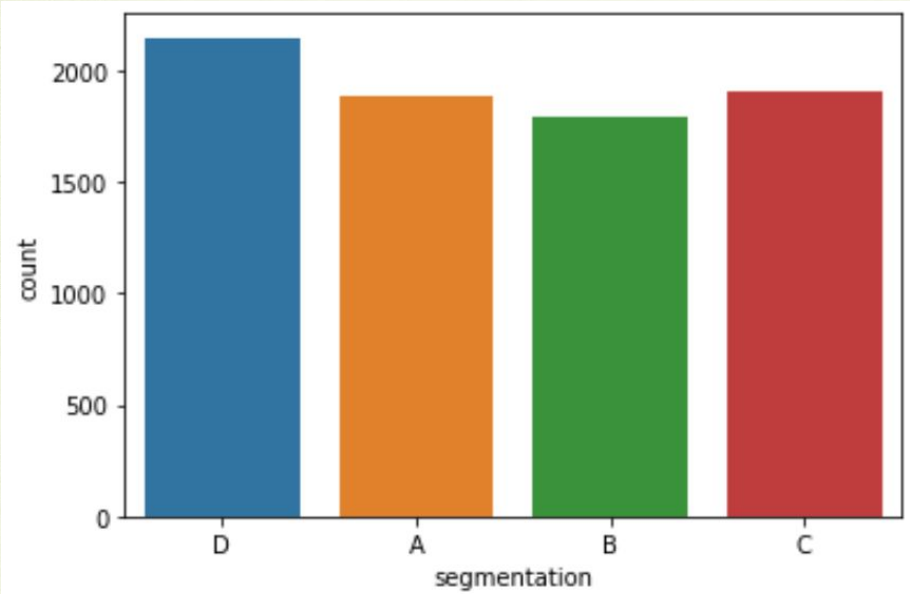
- ever_married, var_1, profession -> drop
- Family_size -> filled with median
- graduated & work_experience
  - If there is profession, work experience will be filled with minimum and graduated will be filled with Yes.

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | id | 8068 non-null | int64 |
| 1 | gender | 8068 non-null | object |
| 2 | ever_married | 7928 non-null | object |
| 3 | age | 8068 non-null | int64 |
| 4 | graduated | 7990 non-null | object |
| 5 | profession | 7944 non-null | object |
| 6 | work_experience | 7239 non-null | float64 |
| 7 | spending_score | 8068 non-null | object |
| 8 | family_size | 7733 non-null | float64 |
| 9 | var_1 | 7992 non-null | object |
| 10 | segmentation | 8068 non-null | object |



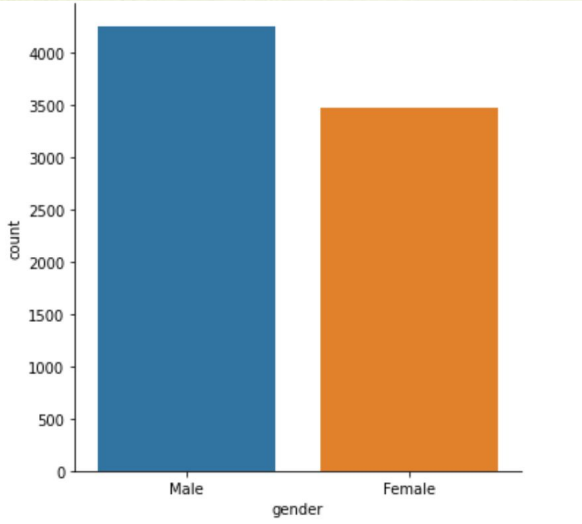
| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | id | 7736 non-null | int64 |
| 1 | gender | 7736 non-null | object |
| 2 | ever_married | 7736 non-null | object |
| 3 | age | 7736 non-null | int64 |
| 4 | graduated | 7736 non-null | object |
| 5 | profession | 7736 non-null | object |
| 6 | work_experience | 7736 non-null | float64 |
| 7 | spending_score | 7736 non-null | object |
| 8 | family_size | 7736 non-null | float64 |
| 9 | var_1 | 7736 non-null | object |
| 10 | segmentation | 7736 non-null | object |

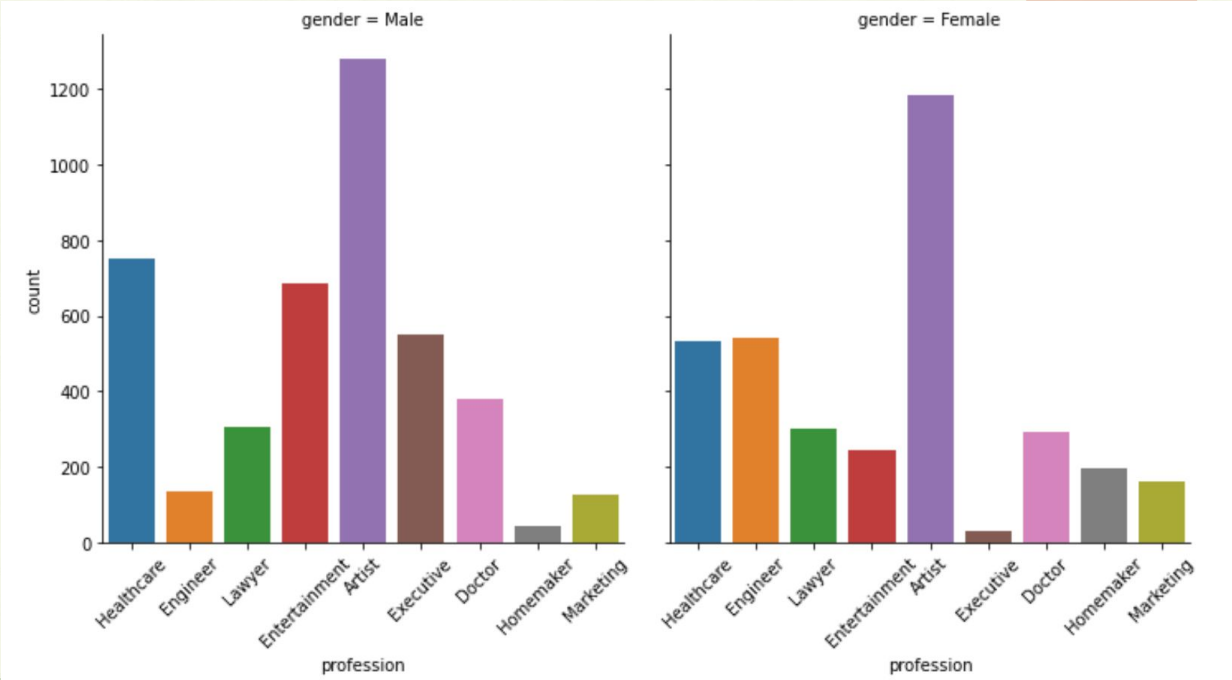
# 2 Data processing and EDA




- 4 customer groups distributed relatively evenly.

# 2 Data processing and EDA




- The gender difference is not big.

- more number of male has a career in healthcare than females.
- More females engineers than males.

# 3 Supervised Learning & Unsupervised learning

## Supervised learning

### Logistic Regression

| | | predicted Segmentation | | | |
|---|---|---|---|---|---|
| | | A | B | C | D |
| actual Segmentation | A | 743 | 201 | 282 | 283 |
| | B | 410 | 342 | 521 | 162 |
| | C | 206 | 193 | 941 | 187 |
| | D | 349 | 84 | 77 | 1207 |

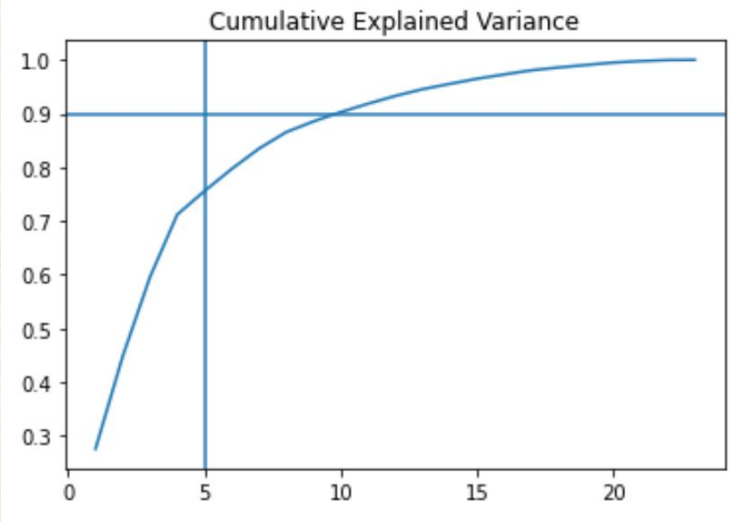
### XGBoost

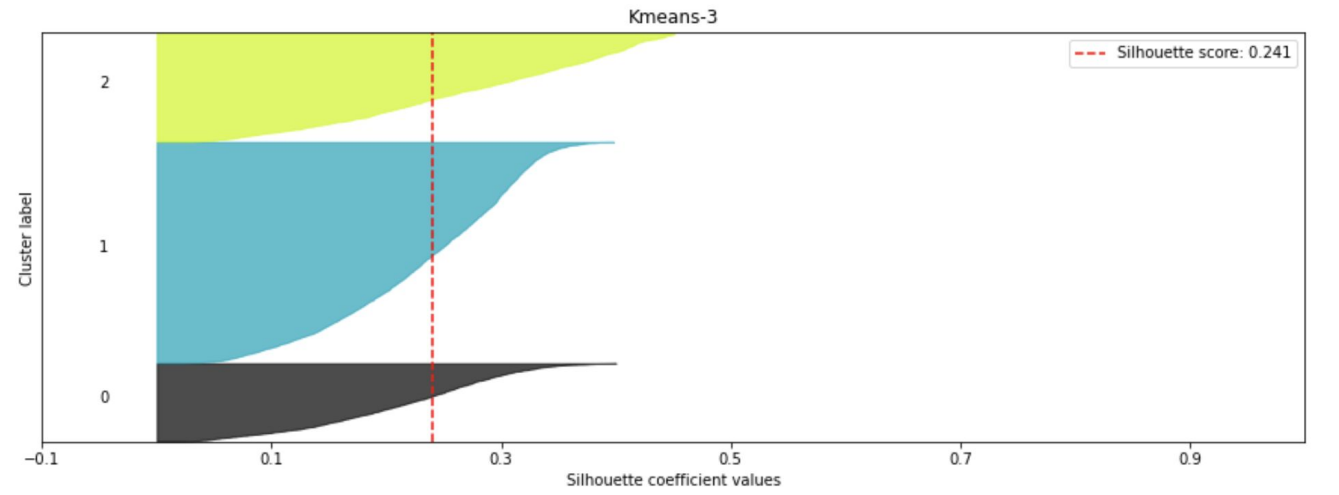
| | | predicted Segmentation | | | |
|---|---|---|---|---|---|
| | | A | B | C | D |
| actual Segmentation | A | 852 | 229 | 169 | 259 |
| | B | 272 | 614 | 394 | 155 |
| | C | 144 | 208 | 991 | 184 |
| | D | 250 | 89 | 38 | 1340 |

# 3 Supervised Learning & Unsupervised learning

## Unsupervised learning

PCA → Kmeans



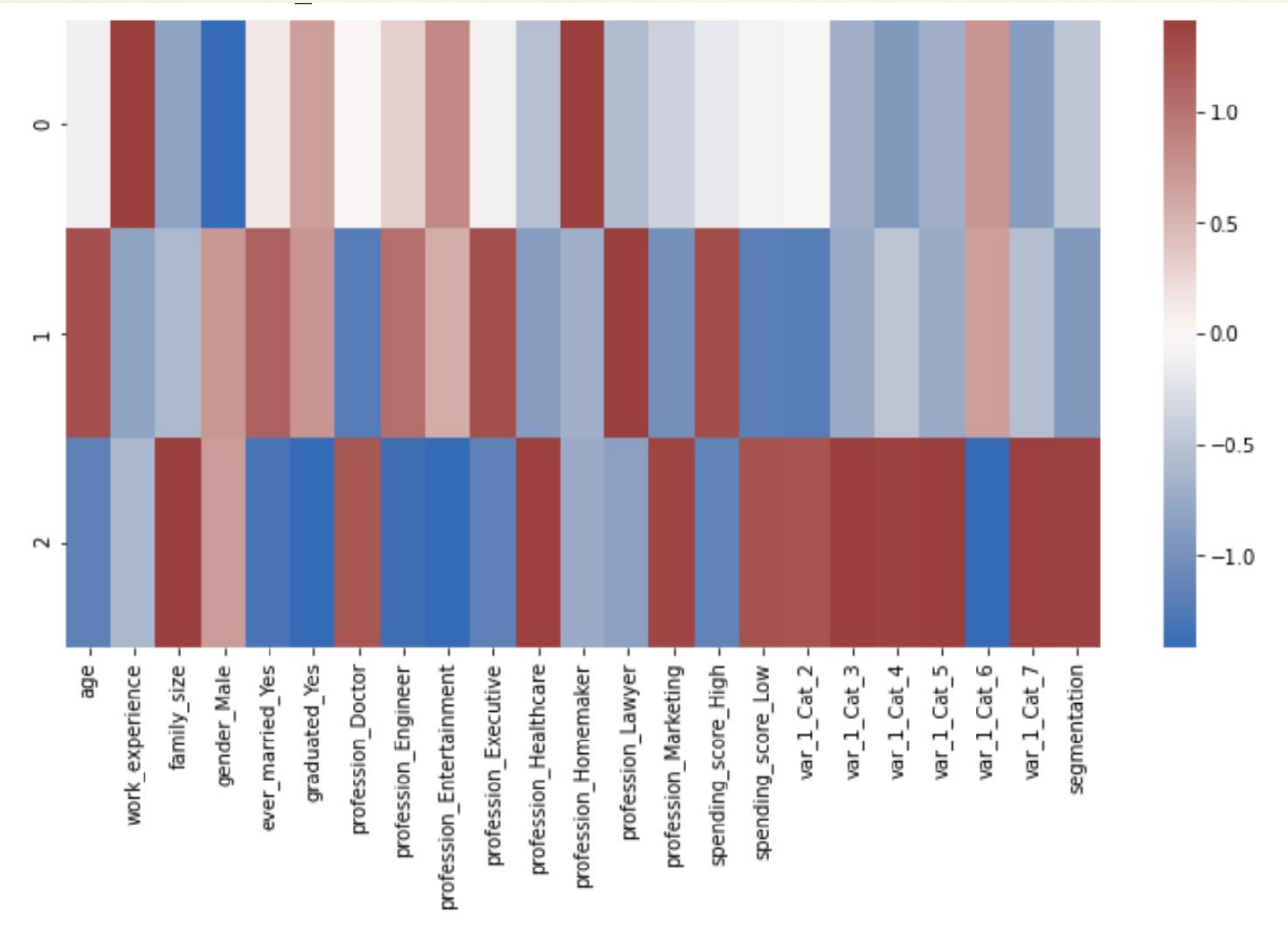




- 10 components will be choosing to reduce features and have enough explained variance.

- 3-Clusters is a good fit.

# 3 Supervised Learning & Unsupervised learning

## Unsupervised learning




- group 0 is highly correlated with female, and also highly correlated with profession of homemaker.
- group 1 is related with age, people who are married, and high paid profession and high spending score.
- group 2 is related with large family size, neither married nor graduated.

# 4 Future research & lessons learned

- Limitation:
  - Dataset is small.
  - Both models are not tuned.

- Future research
  - Try to understand why company make such decision.
  - Try to apply neural network.

- Lessons learned
  - Should keep ethic in mind.
  - Should realize the power of data, and realize the world changing power in data analysis.

# THE END